MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1549 October, 1995
C.B.C.L Paper No. 123

Template Matching: Matched Spatial Filters and beyond

Roberto Brunelli and Tomaso Poggio

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu. The pathname for this publication is: ai-publications/1500-1999/AIM-1549.ps.Z

Abstract

Template matching by means of cross-correlation is common practice in pattern recognition. However, its sensitivity to deformations of the pattern and the broad and unsharp peaks it produces are significant drawbacks. This paper reviews some results on how these shortcomings can be removed. Several techniques (Matched Spatial Filters, Synthetic Discriminant Functions, Principal Components Projections and Reconstruction Residuals) are reviewed and compared on a common task: locating eyes in a database of faces. New variants are also proposed and compared: least squares Discriminant Functions and the combined use of projections on eigenfunctions and the corresponding reconstruction residuals. Finally, approximation networks are introduced in an attempt to improve filter design by the introduction of nonlinearity.

This report describes research done at the Center for Biological and Computational Learning, the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology and at the Istituto per la Ricerca Scientifica e Tecnologica (IRST, Trento, Italy). This research is sponsored by grants from ONR under contract N00014-93-1-0385 and from ARPA-ONR under contract N00014-92-J-1879; and by a grant from the National Science Foundation under contract ASC-9217041 (this award includes funds from ARPA provided under the HPCC program). Support for the A.I. Laboratory's artificial intelligence research is provided by ARPA-ONR contract N00014-91-J-4038. Tomaso Poggio is supported by the Uncas and Helen Whitaker Chair at MIT's Whitaker College. Roberto Brunelli was also supported by IRST.

1 Introduction

The detection and recognition of objects from their images, irrespective of their orientation, scale, and view, is a very important research subject in computer vision, if not computer vision itself. Several techniques have been proposed in the past to solve this challenging problem. In this paper we will focus on a subset of these techniques, those employing the idea of projection to match image patterns. The notion of Matched Spatial Filter (hereafter MSF) is a venerable one with a long history [21]. While by itself it cannot account for invariant recognition, it can be coupled to invariant mappings or signal expansions, and is therefore able to provide invariance to rotation and scaling in the image plane. In order to cope with more general variations of the objects views more sophisticated approaches have to be employed. Among them, the use of Synthetic Discriminant Functions [17, 9, 6, 14, 15, 20, 28, 19, 8, 7, 26, 18, 16] is one of the more promising so far developed. In these paper we will follow a path from MSF, to expansion matching through different variant of SDFs. The first section describes the basic properties of MSF, their optimality and their relation to the probability of misclassification. The generalization of MSF to a linear combination of example images is introduced next. Several shortcomings of the basic approach are outlined and a set of possible solutions is presented in the subsequent section. We discuss a relation of the resulting class of filters to nonorthogonal image expansion. A generalization to projections on multiple directions and the use of the projection residual for pattern matching is then investigated [24, 22, 27, 29, 30, 31]. Finally, a more powerful, non linear framework is introduced in which template matching can be looked at as a problem of function approximation. Network architectures and training strategies are proposed within this new general framework.

2 Matched Spatial Filter

Template matching is extensively used in low-level vision tasks to localize and identify patterns in images. Two methods are commonly used:

1. image subtraction: images are considered as vectors and the norm of their difference is considered as a measure of dissimilarity;

2. correlation: the dot product of two images is considered as a measure of their similarity (it represents the angle between the images when they are suitably normalized and considered as vectors).

When the images are normalized to have zero average and unit norm, the two approaches give the same result. The usual implementation of the above methods relies on the euclidean distance. Other distances can be used and some of them have better properties such as increased *robustness* to noise and minor deformations [4]. The next sections are mainly concerned with the correlation approach. The idea of image subtraction is introduced again in the more general nonlinear framework.

2.1 Optimality

One of the reasons for which template matching by correlation is commonly used is that correlation can be shown to be the optimal (according to a particular criterion) linear operation by which a deterministic reference function can be extracted from additive white Gaussian noise [21]. Let the detected signal be:

$$g(x) = \phi(x) + \lambda(x) \tag{1}$$

where $\phi(x)$ is the original signal and $\lambda(x)$ is noise with power spectrum $S(\omega)$. The noise is assumed to be wide-sense stationary with zero average so that:

$$\begin{aligned} E\{\lambda(x)\} &= 0 \\ E\{\lambda(x+\alpha)\lambda(x)\} &= R(\alpha) \end{aligned}$$

We assume that $\phi(x)$ is known and we want to establish its presence and location. To do so we apply to the process $g(x)$ a linear filter with impulse response $h(x)$ and system function $H(\omega)$. The resulting output is:

$$\begin{aligned} z(x) &= g(x) * h(x) = \int_{-\infty}^{\infty} g(x-\alpha)h(\alpha)d\alpha && (2) \\ &= z_\phi(x) + z_\lambda(x) && (3) \end{aligned}$$

Using the convolution theorem for the Fourier transform we have that:

$$\begin{aligned} z_\phi(x) &= \int_{-\infty}^{\infty} \phi(x-\alpha)h(\alpha)d\alpha \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi(\omega)H(\omega)e^{i\omega x}d\omega \end{aligned}$$

We want to find $H(\omega)$ so as to maximize the following signal to noise ratio (SNR):

$$r^2 = \frac{|z_\phi(x_0)|^2}{E\{z^2{}_\lambda(x_0)\}} \tag{4}$$

where x_0 is the location of the signal. The SNR represents the ratio of the filter responses at the uncorrupted signal and at the noise. It is defined at the true location of the signal (usually the correlation peak) therefore not taking into account the off-peak response of the filter.

Two cases of particular interest are those of white and colored noise:

White Noise This type of noise is defined by the following condition

$$S(\omega) = S_0$$

which corresponds to a flat energy spectrum. The Schwartz inequality states that

$$|\int_a^b f(x)g(x)dx|^2 \leq \int_a^b |f(x)|^2 \int_a^b |g(x)|^2$$

and the equality hold iff $f(x) = k\tilde{g}(x)$ (we use $\tilde{}$ to represent complex conjugation). This implies the following bound for the signal to noise ratio r:

$$r^2 \leq \frac{\int |\Phi(\omega)e^{i\omega x_0}|^2 d\omega \int |H(\omega)|^2 d\omega}{2\pi S_0 \int |H(\omega)|^2 d\omega}$$

and then

$$r^2 \leq \frac{E_\Phi}{S_0}$$

where

$$E_\Phi = \frac{1}{2\pi}\int |\Phi(\omega)|^2 d\omega$$

represents the energy of the signal. From the Schwartz inequality the equality holds only if

$$H(\omega) = k\tilde{\Phi}(\omega)e^{-i\omega x_0}$$

The spatial domain version of the filter is simply the mirror image of the signal:

$$h(x) = k\phi(x_0 - x)$$

which implies that the convolution of the signal with the filter can be expressed as the cross-correlation with the signal (hence the name Matched Spatial Filter).

Colored Noise If the noise has a non flat spectrum $S(\omega)$ it is said to be colored. In this case the following holds:

$$\begin{aligned} 2\pi z_\phi(x_0) &= \int \Phi(\omega)H(\omega)e^{i\omega x}d\omega \\ |2\pi z_\phi(x_0)|^2 &= |\int \frac{\Phi(\omega)}{\sqrt{S(\omega)}}\sqrt{S(\omega)}H(\omega)e^{i\omega x}d\omega|^2 \\ &\leq \int \frac{|\Phi(\omega)e^{i\omega x}|^2}{S(\omega)}d\omega \int S(\omega)|H(\omega)|^2 d\omega \end{aligned}$$

hence

$$r^2 \leq \frac{1}{2\pi}\int \frac{|\Phi(\omega)e^{i\omega x}|^2}{S(\omega)}d\omega$$

with equality holding only when

$$\sqrt{S(\omega)}H(\omega) = k\frac{\tilde{\Phi}e^{-i\omega x_0}}{\sqrt{S(\omega)}}$$

The main consequence of the color of noise is that the optimal filter corresponds to a modified version of the signal

$$H(\omega) = k\frac{\tilde{\Phi}e^{-i\omega x_0}}{S(\omega)}$$

which emphasizes the frequencies where the energy of the noise is smaller. The optimal filter can also be considered as a cascade of a whitening filter $S^{-1/2}(\omega)$ and the usual filter based on the transformed signal.

In the spatial domain, correlation amounts to projecting the signal $g(x)$ onto the available *template* $\phi(x)$. If the norm of the projected signal is not equal to that of the template, the value of the projection can be meaningless as the projection value can be large without implying that the two vectors are close in any reasonable sense. The solution is to compute the projection using normalized vectors. In particular, if versors are used, computing the projection amounts to computing the cosine of the angle formed by the two vectors, which is an effective measure of similarity. In vision tasks, vector normalization corresponds to adjusting the intensity scale so that the corresponding distribution has a given variance. Another useful normalization is to set the average value of the vector coordinates to zero. This operation corresponds to setting the average of the intensity distribution for images. These normalization are particularly useful when modern cameras are used, as they usually operate with automatic gain level (acting on the scale of the intensity) and black level adjustment (acting as on offset on the intensity distribution).

2.2 Distorted Templates

The previous analysis was focused on the detection of a deterministic signal corrupted by noise. An interesting extension is the detection of a signal belonging to a given distribution of signals [17]. As an example, consider the problem of locating the eyes in a face image. We do not know who's face it is so that we cannot select the corresponding signal (the eyes of that person). A whole set of different eyes could be available, possibly including the correct ones.

Let $\{\phi(x)\}$ denote the class of signals to be detected. We want to find the filter h which maximizes the SNR r^2 over the class of signals $\{\phi(x)\}$. The input signal $\phi(x)$ can be modeled as a sample realization of the stochastic process $\{\phi(x)\}$. The ensemble-average correlation function of the stochastic process is defined by

$$K_{\phi\phi}(x, y) = E_\phi\{\phi(x)\phi(y)\} \tag{5}$$

and represents the average over the ensemble of signals (and not over the coordinates of a signal). What we want to maximize is the ensemble average of the signal to noise ratio:

$$E_\phi\{r^2\} = \frac{E\{|z_\phi(x_0)|^2\}}{E\{z^2{}_\lambda(x_0)\}} \tag{6}$$

Assume, without loss of generality, that $x_0 = 0$. The average SNR can then be rewritten as:

$$E_\phi\{r^2\} = \frac{\int\int h(-x)h(-y)K_{\phi\phi}(x,y)dxdy}{\int\int h(-x)h(-y)K_{\lambda\lambda}(x,y)dxdy} \tag{7}$$

where the ensemble autocorrelation function of the signal and noise have been used. The autocorrelation function of the white noise is proportional to a Dirac delta function:

$$K_{\lambda\lambda}(x, y) = N\delta(x - y) \tag{8}$$

so that the average signal to noise ratio can be rewritten as:

$$E_\phi\{r^2\} = \frac{\int\int h(-x)h(-y)K_{\phi\phi}(x,y)dxdy}{N\int h(-x)^2 dx} \tag{9}$$

Pre-whitening operators can be applied as preprocessing functions when the assumption of white noise does not hold. The denominator of the RHS in eqn. (9) represents the energy of the filter and we can require it to be 1:

$$\int h(-x)^2 dx = 1 \tag{10}$$

To optimize eqn. (9) we must maximize the numerator subject to the energy constraint of the filter. The ensemble auto-correlation function can be expressed in terms

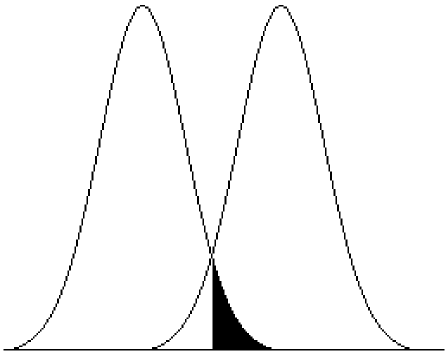

Figure 1: The probability of error, represented by the shaded area, when the distributions are Gaussian with the same covariance.

of the orthonormal eigenfunctions of the integral kernel $K_{\phi\phi}(x, y)$

$$K_{\phi\phi}(x, y) = \sum_i \lambda_i \psi_i(x) \psi_i(y) \tag{11}$$

where the λ_i are the corresponding eigenvalues. The filter function h can also be expanded in the same basis

$$h(-x) = \sum_i \omega_i \psi_i(x) \tag{12}$$

Using the inner product notation and the orthonormality of the $\psi_i(x)$ we can state the optimization problem as finding

$$arg \max_{\sum_i \omega^2{}_i = 1} \sum_i \lambda_i (h \cdot \psi_i)^2 \tag{13}$$

If we order the eigenvalues so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq \ldots$, we have

$$N \cdot E_\phi\{r^2\} = \sum_i \lambda_i (h \cdot \psi_i)^2 = \sum_i \lambda_i \omega_i^2 \leq \lambda_1 \sum_i \omega_i^2 = \lambda_1 \tag{14}$$

and the maximum value is achieved when the filter function is taken to be the dominant eigenvector.

2.3 Signal-to-noise ratio and classification error

Several performance metrics are available for correlation filters that describe attributes of the correlation plane. The signal to noise ratio (SNR) is just one of them. Other useful quantities are the peak-to-correlation energy, the location of the correlation peak and the invariance to distortion. As correlation is typically used to locate and discriminate objects, another important measure of a filter's performance is how well it discriminates between different classes of objects. The simplest case is given by the discrimination between the signal and the noise. In this section we will show [16, 12] that for the classical matched filter maximizing the SNR is equivalent to minimizing the probability of classification error P_e when the underlying probability distribution functions (PDFs) are Gaussians.

The classifier which minimizes the probability of error is the Bayes classifier. For two normal distributions,

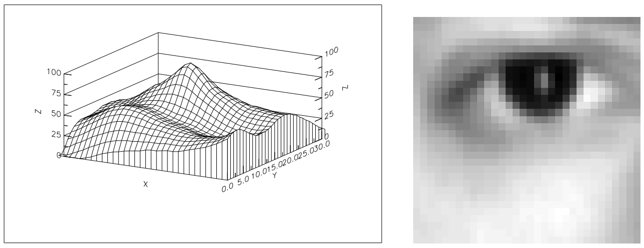

Figure 2: The cross-correlation of the template reported on the right. Note the diffuse shape of the peak that makes its localization difficult

the Bayes decision rule can be expressed as a quadratic function of the observation vector $\boldsymbol{x}$ as

$$\begin{aligned} \frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_A)^T \Sigma_A^{-1} (\boldsymbol{x} - \boldsymbol{m}_A) \quad & - \\ \frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_B)^T \Sigma_B^{-1} (\boldsymbol{x} - \boldsymbol{m}_B) \quad & + \\ \frac{1}{2} \ln \frac{|\Sigma_A|}{|\Sigma_B|} \begin{array}{c} A \\ > \\ < \\ B \end{array} \ln \frac{P_A}{P_B} & \end{aligned} \tag{15}$$

where $\boldsymbol{m}_A, \boldsymbol{m}_B$ are the distribution means, Σ_A, Σ_B the covariance matrices and P_A, P_B the occurence probabilities.

Let us consider two classes: a deterministic signal $\boldsymbol{\phi}$ corrupted with white Gaussian noise as class A and the noise itself as class B. In this case $\boldsymbol{m}_A = \boldsymbol{\phi}$, $\boldsymbol{m}_B = \vec{0}$ and $\Sigma_A = \Sigma_B = I$. This means that the components of the signal $\boldsymbol{\phi}$ are uncorrelated and have unit variance. If we further assume that the *a priori* probabilities of occurence of these classes are equal, the probability of error (see also Figure 1) is given by:

$$P_e = \frac{1}{\sqrt{2\pi}} \int_\eta^\infty exp(-u^2/2) du \tag{16}$$

where $\eta = \frac{1}{2}\xi^{1/2}$, with ξ being the Mahalanobis distance between the PDFs of the two classes:

$$\xi = (\boldsymbol{m}_A - \boldsymbol{m}_B)^T I (\boldsymbol{m}_A - \boldsymbol{m}_B) = \boldsymbol{\phi}^T \boldsymbol{\phi} \tag{17}$$

and the Bayes decision rule simplifies to:

$$\boldsymbol{x} \in A \ \text{ if } \boldsymbol{\phi}^T \boldsymbol{x} > \tfrac{1}{2}\xi \tag{18}$$

$$\boldsymbol{x} \in B \ \text{ if } \boldsymbol{\phi}^T \boldsymbol{x} \leq \tfrac{1}{2}\xi \tag{19}$$

The input vector $\boldsymbol{x}$ is then classified as signal or noise depending on the value of the correlation with the uncorrupted signal. We have already shown that correlation with the signal maximizes the signal to noise ratio, so when the noise distribution is Gaussian maximizing the SNR is equivalent to minimizing the classification error probability. When the noise is not white, the signal can be transformed by applying a whitening transformation A:

$$A^T \Sigma A = I \tag{20}$$

and the previous reasoning can be applied.

3 Synthetic Discriminant Functions

While correlators are optimal for the recognition of patterns in the presence of white noise they have three major limitations: the output of the correlation peak degrades rapidly with geometric image distortions, the peak is often broad (see Figure 2), making its detection difficult, and they cannot be used for multiclass pattern recognition. It has been noted that one can obtain better performance from a multiple correlator (i.e. one computing the correlation with several templates) by forming a linear combination of the resulting outputs instead of, for example, taking the maximum value [10, 11]. The filter synthesis technique known as Synthetic Discriminant Functions (hereafter SDF) starts from this observation and builds a filter as a linear combination of MSFs for different patterns [9, 6]. The coefficients of the linear combination are chosen to satisfy a set of constraints on the filter output, requiring a given value for each of the patterns used in the filter synthesis. By forcing the filter output to different values for different patterns, multiclass pattern recognition can be achieved. Let $\{\phi_i(x)\}_{i=1,\ldots,n}$ be a set of (linearly independent) images and $\boldsymbol{u} = \{u_1, \ldots, u_n\}^T$ be a vector representing the required output of the filter for each of the images:

$$\phi_i \otimes h = u_i \tag{21}$$

where $\otimes$ represents correlation (not convolution). The filter h can be expressed as a linear combination of the images ϕ_i:

$$h(x) = \sum_{i=1,\ldots,n} b_i \phi_i(x) \tag{22}$$

as any additional contribution from the space orthogonal to the images would yield a zero contribution when correlating with the image set. If we denote by $\boldsymbol{X}$ the matrix whose columns represent the images (represented as vectors by concatenating their rows), by enforcing the constraints we obtain the following set of equations:

$$\boldsymbol{b} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{u} \tag{23}$$

which can be solved as the images are linearly independent. The resulting filter is appropriate for pattern recognition applications in which the input object can be a member of several classes and different distorted versions of the same object (or different objects) can be expected within each class. If M is the number of classes, n_i is the number of different pattern within each class i, N the overall number of patterns, M filters can be built by solving

$$\boldsymbol{b}_i = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{\delta}_i \quad i = 1, \ldots, M \tag{24}$$

where

$$\delta_{ik} = \begin{cases} 1 & \sum_{j=1}^{i-1} n_j < k < \sum_{j=1}^{i} n_j \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

$k = 1, \ldots, N$ and image ϕ_k belongs to class i if $\delta_{ik} = 1$. Discrimination of different classes can be obtained also using a single filter and imposing different output values. However the performance of such a filter is expected to be inferior to that of a set of class specific filters due to the high number of constraints imposed on the filter outputs [9]. While this approach makes it easy to obtain predefined values on a given set of patterns it does not allow to control the off-peak filter response. This can prevent reliable classification when the number of constraints becomes large.

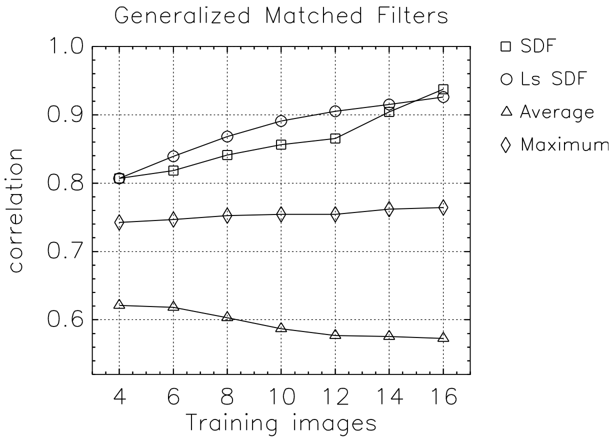

Figure 3: An increasing portion of a set of 30 eyes images was used to build a SDF, an average MSF or a set of prototype MSFs from which the highest response was extracted. Our new least square SDF uses four building templates. The plot report the average responses over a disjoint 30 image test set. Note that the lower values of MSFs are due to the fact that their response is not *scaled* to obtained a predefined value as opposed to SDFs whose output is constrained to be 1, and to approximate 1 for ls SDFs.

The effect of filter *clutter* can also appear in the construction of a filter giving a fixed response over a set of images belonging to the same class (the Equal Correlation Filter introduced in [9]).

In order to minimize this problem we propose a new variant of SDFs: least quares SDFs. These filters are computed using only a subset of the training images[1] and the coefficient of the linear combination are chosen to minimize the square error of the filter output on all of the available images. In this case the matrix $\boldsymbol{R} = \boldsymbol{X}^T \boldsymbol{X}$ is rectangular and the estimate of the $\boldsymbol{b}$ relies on the computation of the pseudoinverse of $\boldsymbol{R}$:

$$\boldsymbol{R}^\dagger = (\boldsymbol{R}^T \boldsymbol{R})^{-1} \boldsymbol{R}^T \tag{26}$$

The dimension of the matrix to be inverted is $n \times n$ where n represents the number of images used to build the filter and not the (greater) number of training images. By using a reduced number of building templates the problem of filter cluttering is reduced. A different use of least square estimation for filter synthesis can be found in [6] where it is coupled to Karhunen-Loeve expansion for the construction of correlation SDFs.

[1] The subset of training images can be chosen in a variety of ways. In the reported experiments they were chosen at random. Another possibility is that of clustering the available images, the number of clusters being equal to the number of images used in filter synthesis.

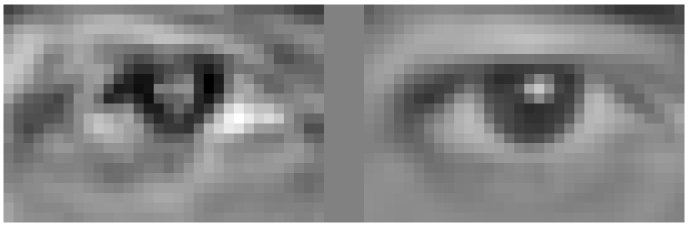

Figure 4: The MSFs resulting from using 20 building images in the SDF (left) and 2 in the least square SDF (right) when using the same set of training images. The difference in contrast of the two images reflect the magnitude of the MSFs. The performance of the two filters was similar.

The results for a sample application are reported in Figure 3. Note that by using a least square estimate a good performance can be achieved using a small number of templates. This has a major influence on the appearance of the resulting MSF as can be seen in Figure 4.

Another variant is to use *symbolic encoded filters* [9]. In this case a set of k filters is built whose outputs are 0 or 1 and can be used to encode the different patterns using a binary code. In order to use the filter for classification, the outputs are thresholded and the resulting binary number is used to index the pattern class.

Synthesis of the MSF from a projection SDF algorithm can achieve distortion invariance and retain shift invariance. However, the resulting filter cannot prevent large sidelobe levels from occurring in the correlation plane for the case of false (or true) targets. The next section will detail the construction of filters which guarantee controlled sharp peaks and good noise immunity.

4 Advanced SDFs

The signal to noise ratio maximized by the MSF is limited to the correlation peak: it does not take into account the off-peak response and the resulting filters often exhibit a sustained response well apart from the location of the central peak. This effect is usually amplified in the case of SDF when many constraints are imposed on the filter output. In order to locate the correlation peak reliably, it should be very localized [14]. However, it can be expected that the greater the localization of the filter response (approaching a δ function) the more sensitive the filter to slight deviations from the patterns used in its synthesis. This suggests that the best response of the filter should not really be a δ function, but some shape, like a Gaussian, whose dispersion can be tuned to the characteristics of the pattern space. In this section we will review the synthesis of such filters in the frequency domain [26].

Let us assume for the moment that there is no noise. The correlation of the i-th pattern with the filter h is represented by

$$z_i(n) = \phi_i(n) \otimes h(n) \quad n = 0, \ldots, d-1 \tag{27}$$

where d is the dimension of the patterns. In the following capital letters are used to denote the Fourier transformed quantities. The filter is also required to produce an output u_i for each training image:

$$z_i(0) = u_i \tag{28}$$

which can be rewritten in the Fourier domain as:

$$\boldsymbol{H}^+ \boldsymbol{X} = d\boldsymbol{u} \tag{29}$$

where $^+$ denotes complex conjugate transpose.Using Parseval's theorem, the energy of the i-th *circulant* correlation plane is given by:

$$E_i = \sum_{n=0}^{d-1} |z_i(n)|^2 = \frac{1}{d}\sum_{k=0}^{d-1} |Z_i(k)|^2 = \frac{1}{d}\sum_{k=0}^{d-1} |H(k)|^2 |\Phi_i(k)|^2 \tag{30}$$

When the signal is perturbed with noise the output of the filter will also be corrupted:

$$z_i(0) = \phi_i(0) \otimes h(0) + \lambda(0) \otimes h(0) \tag{31}$$

Under the assumption of zero-mean noise, the variance of the filter output due to noise is:

$$E_N = \frac{1}{d}\sum_{k=0}^{d-1} |H(k)|^2 S(k) \tag{32}$$

where $S(k)$ is the noise spectral energy. What we would like is a filter whose average correlation energy over the different training images and noise is as low as possible while meeting the constraints on the filter outputs. A first choice is to minimize:

$$\begin{aligned} E &= \sum_i (E_i + E_N) && (33) \\ &= \frac{1}{d}\sum_i \sum_k |H(k)|^2 (|\Phi_i(k)|^2 + S(k)) && (34) \end{aligned}$$

subject to the constraints of eqn. (29). However, minimizing the average energy (or filter variance due to noise) does not minimize each term, corresponding to a particular correlation energy (or noise variance). A more stringent bound can be obtained by considering the spectral envelope of the different terms in eqn. (34):

$$E = \sum_k |H(k)|^2 \max(\Phi_1(k)|^2, \ldots, |\Phi_N(k)|^2, S(k)) \tag{35}$$

If we introduce the diagonal matrix $T_{kk} = N \max(|\Phi_1(k)|^2, \ldots, |\Phi_N(k)|^2, S(k))$ the filter synthesis can be summarized as minimizing

$$E = \boldsymbol{H}^+ \boldsymbol{T} \boldsymbol{H} \tag{36}$$

subject to

$$\boldsymbol{H}^+ \boldsymbol{X} = d\boldsymbol{u} \tag{37}$$

This problem can be solved [20] using the technique of Lagrange multipliers to minimize the function:

$$\mathcal{E} = \boldsymbol{H}^+ \boldsymbol{T} \boldsymbol{H} - 2\sum_{i=1}^{N} \lambda_i (\boldsymbol{H}^+ \boldsymbol{X}_i - du_i) \tag{38}$$

where $\lambda_1, \ldots, \lambda_N$ are the parameters introduced to satisfy the constrained minimization. By zeroing the gradient of $\mathcal{E}$ with respect to $\boldsymbol{H}$ we can express $\boldsymbol{H}$ as a

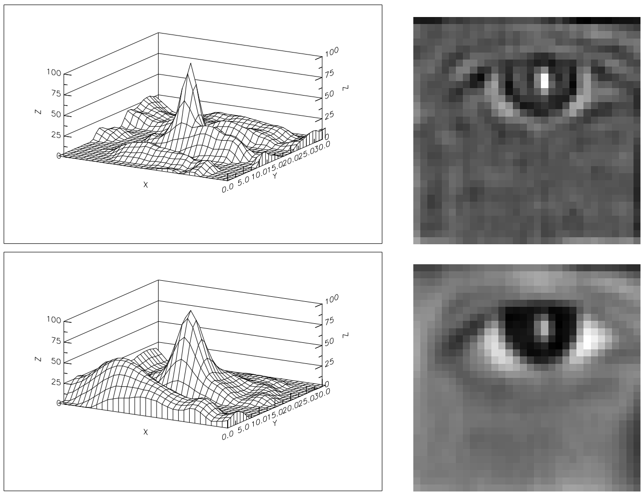

Figure 5: Using an increasing amount of added white noise the emphasis given to the high frequency is reduced and the resulting filter response approaches that of the standard MSF.

function of $\boldsymbol{T}$ and of $\Lambda = \{\lambda_1, \ldots, \lambda_N\}$. By substitution into eqn. (37) the following solution is found:

$$\boldsymbol{H} = \boldsymbol{T}^{-1}\boldsymbol{X}(\boldsymbol{X}^{+}\boldsymbol{T}^{-1}\boldsymbol{X})^{-1}\boldsymbol{u} \tag{39}$$

The use of the spectral envelope has the effect of reducing the emphasis given by the filter to the high frequency content of the signal, thereby improving intraclass performance. It is important to note that the resulting filter can be seen as a cascade of a *whitening* filter $\boldsymbol{T}^{-1/2}$ and a conventional SDF based on the transformed data. Note that in this case the whitened spectrum is the envelope of the spectra of the real noise and of the training images. A least square approach may again be preferred to cope with a large number of examples. In this case all available images are used to estimate $\boldsymbol{T}$ but only a subset of them is used to build the corresponding SDF. Experiments have been reported using a white noise of tunable energy α to model the intraclass variability [26]

$$E = \sum_k |H(k)|^2 \max(|\Phi_1(k)|^2, \ldots, |\Phi_N(k)|^2, \alpha) \tag{40}$$

Adding white noise limits the emphasis given to high frequencies, reducing the sharpness of the correlation peak and increasing the tolerance to small variations of the templates (see Figures 5 and 6). A comparison of different filters is reported in Figure 7. The effects of non linear processing emphasizing the high frequencies to obtain a sharp correlation peak is reported in Figure 8.

Another way of controlling the intraclass performance is that of modeling the correlation peak shape [8, 18]. As already mentioned, the shape of the correlation peak is expected to be important both for its detection and for the requirements imposed on the filter which can impair its ability to correlate well with patterns even slightly different from the ones used in the training. Let us denote with $F(k)$ the required shape of the correlation peak. The shape of the peak can be constrained by minimizing the squared deviations of its output from the required shape F:

$$E_S = \sum_{i=1}^{N}\sum_{k=1}^{d} |H(k)^*\Phi_i(k) - F(k)|^2 \tag{41}$$

where, for instance, $f(x) = exp(-x^2/2\sigma^2)$ is a Gaussian amplitude function. By switching to matrix notation, the resulting energy can be expressed as:

$$E_S = \boldsymbol{H}^{+}\boldsymbol{D}\boldsymbol{H} + \boldsymbol{F}^{+}\boldsymbol{F} - \boldsymbol{H}^{+}\boldsymbol{A}\boldsymbol{F} - \boldsymbol{F}^{+}\boldsymbol{A}^{+}\boldsymbol{H} \tag{42}$$

where $\boldsymbol{A}$ is a diagonal matrix whose elements are the sum of the components of $\boldsymbol{\Phi}_i$ and $\boldsymbol{D}$ is a diagonal matrix whose elements are the sum of the squares the components of $\boldsymbol{\Phi}_i$. The first term in the RHS of eqn. (42) corresponds to the average correlation energy of the different patterns (see eqn. (30)). We suggest the use of the spectral envelope $\boldsymbol{T}$ instead of $\vec{D}$, employed in the original approach, thereby minimizing the following energy

$$E'_S = \boldsymbol{H}^{+}\boldsymbol{T}\boldsymbol{H} + \boldsymbol{F}^{+}\boldsymbol{F} - \boldsymbol{H}^{+}\boldsymbol{A}\boldsymbol{F} - \boldsymbol{F}^{+}\boldsymbol{A}^{+}\boldsymbol{H} > E_S \tag{43}$$

The minimization of E_S subject to the constraints of eqn. (29) can be done again using the Lagrange multiplier and is found to be:

$$\begin{aligned} \boldsymbol{H} \; = \; & \boldsymbol{T}^{-1}\boldsymbol{X}(\boldsymbol{X}^{+}\boldsymbol{T}^{-1}\boldsymbol{X})^{-1}d\boldsymbol{u} \\ & + \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{F} - \boldsymbol{T}^{-1}\boldsymbol{X}(\boldsymbol{X}^{+}\boldsymbol{T}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{+}\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{F} \end{aligned} \tag{44}$$

These filters provide a controlled, sharp correlation peak subject to the constraints on the filter output, the required correlation peak shape and the reduce variance to the noise. In our experiments the Fourier domain was used to compute the *whitening* filters. They were then transformed to the spatial domain where a standard correlation was computed after their application. An approach using only computations in the space domain can be found in [28].

5 Nonorthogonal Image Expansion and SDF

In this section we review an alternative way of looking at the problem of obtaining sharp correlation peaks, namely the use of nonorthogonal image expansion [2, 3]. Matching by expansion is based on expanding the signal with respect to basis functions (BFs) that are all translated versions of the template. Such an expansion is feasible if the BFs are linearly independent and complete. It can be proven that self-similar BFs of compact support are independent and complete under very weak conditions. Suppose one wants to estimate the discrete d-dimensional signal $g(x)$ by a linear combination of basis functions $\phi_i(x)$:

$$g'(x) = \sum_{i=1}^{d} c_i \phi_i(x) \tag{45}$$

where $\phi_i(x)$ now represents the i-th circulated translation of ϕ. The coefficients are estimated by minimizing

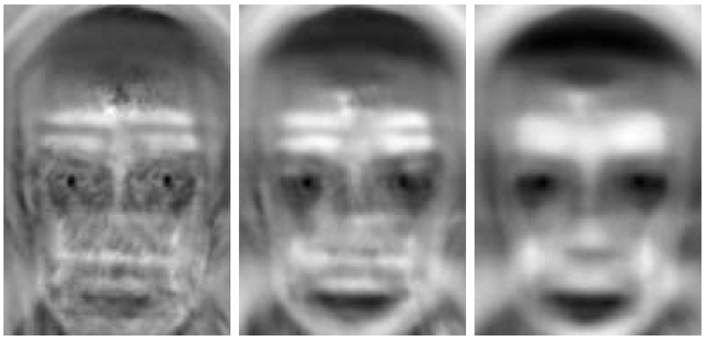

Figure 6: The output of the correlation with an SDF computed using the spectral envelope of 10 training images and different amounts of white noise (left: $\alpha = 1$, middle $\alpha = 5$) compared to the output of normalized cross-correlation using one of the images used to build the SDF but without any spectral enhancement. The darker the image the higher the corresponding value. Note that an increased amount of white noise improves the response of the filter.

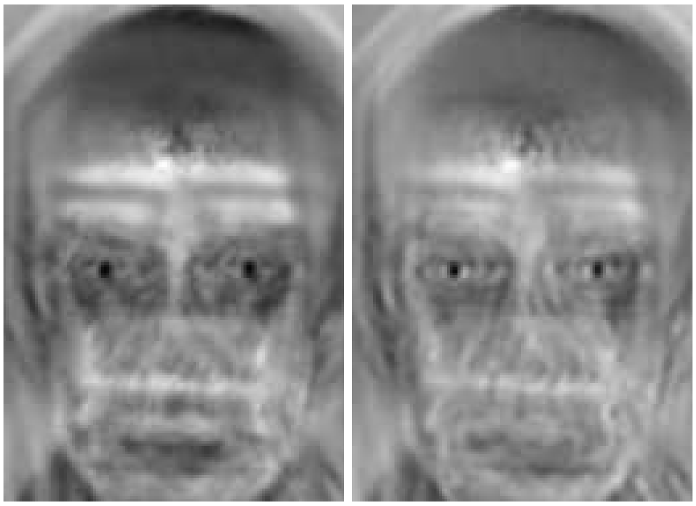

Figure 7: The output of the correlation with an SDF computed using the spectral envelope of 20 training images as whitening preprocessing. Left: the normal SDF (20 examples). Right: a least square SDF with 6 templates (20 examples). The darker the image the higher the corresponding value. The least square SDF exhibits a sharper response using the same whitening filter.

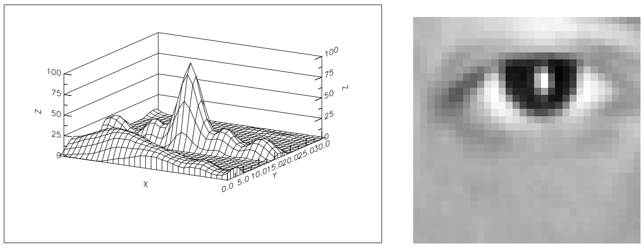

Figure 8: Non linear processing can be employed. The figure represent the result of preprocessing the image to extract the local image contrast (intensity value over the average value in a small neighborhood). This kind of preprocessing emphasizes high frequencies and results in a sharp correlation peak.

the square error of the approximation $|g - g'|^2$. The approximation error is orthogonal to the basis functions so that the following system of equations must be solved:

$$\begin{aligned} \sum_{j=1}^{d} < \boldsymbol{\phi}_1, \boldsymbol{\phi}_j > c_j &= < \boldsymbol{g}, \boldsymbol{\phi}_1 > \qquad (46) \\ &\cdots \\ \sum_{j=1}^{d} < \boldsymbol{\phi}_d, \boldsymbol{\phi}_j > c_j &= < \boldsymbol{g}, \boldsymbol{\phi}_d > \end{aligned}$$

If the set of basis functions is linearly independent the equations give a unique solution for the expansion coefficients. If we consider the advanced SDF for the case of no-noise, single training image and working in the spatial domain [28], we have that the corresponding filter can be expressed as:

$$\boldsymbol{h} = ([\boldsymbol{\phi}]^T [\boldsymbol{\phi}])^{-1} \boldsymbol{\phi} \qquad (47)$$

where the columns of matrix $[\cdot]$ are the circulated basis functions ϕ_i. The output of the correlation is then given by:

$$[\boldsymbol{\phi}]\boldsymbol{h} = \boldsymbol{c} = [\boldsymbol{\phi}^T]^{-1} \boldsymbol{\phi} \qquad (48)$$

The solution of the system (47) can be expressed as:

$$\boldsymbol{c} = [\boldsymbol{\phi}^T]^{-1} \boldsymbol{\phi} \qquad (49)$$

which is clearly the same. In the case of no noise the resulting expansion is $\boldsymbol{c} = (0, \ldots, 0, 1, 0, \ldots 0)$ with a single 1 at the location of the signal. The idea of expansion matching is also closely related to correlation SDFs [6] where multiple shifted templates were explicitly used to shape the correlation peak. Let us consider a set of templates obtained by shifting the original pattern (possibly with circulation) on the regular grid defined by the image coordinates. We can require that the correlation value of the original pattern with its shifted versions be 1 when there is no shift and 0 for every non null shift. This corresponds to a filter whose response is given by $\boldsymbol{c} = (0, \ldots, 0, 1, 0, \ldots 0)$ as previously described.

6 Other projection approaches

The whole idea of projection Synthetic Discriminant Functions is to find a direction onto which the pro-

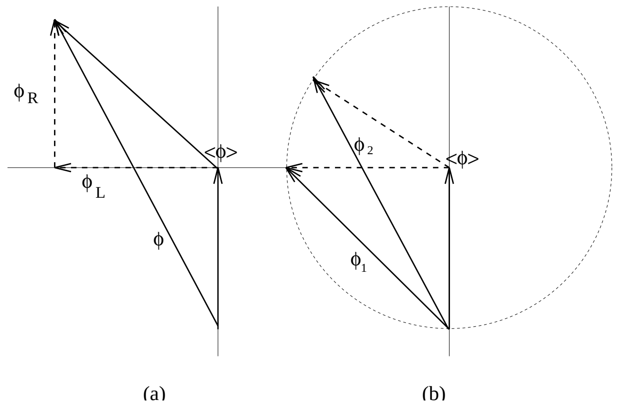

Figure 9: Computing the distance from linear subspace (a) versus computing the distance from a single prototype (b). In drawing (a), vector $< \boldsymbol{\phi} >$ represents the average pattern and the horizontal line on which $\boldsymbol{\phi}_L$ lies represents the linear subspace. $\boldsymbol{\phi}_L$ is the projection of pattern $\boldsymbol{\phi}$ on the linear space and $\boldsymbol{\phi}_R$ is the projection residual. Drawing (b) shows two vectors $\vec{\phi}_1$ and $\boldsymbol{\phi}_2$ with the same distance from the average pattern $\boldsymbol{A}$ but different distances from the pattern space.

jections of the different signals have predefined values. A typical image with 256 × 256 pixels is projected, for recognition purposes, onto a single direction in this high dimensional space. Another approach is to project the signal to be recognized onto a linear subspace [32, 27, 22, 29, 30, 31]. Let us first assume that the patterns of each of the classes to be discriminated belongs to different linear subspaces. For each class it is then possible to determine an orthogonal transformation which diagonalizes the covariance matrix. The elements of the transformed basis are the eigenvectors of the covariance matrix and are called principal components. They can be sorted by decreasing contribution to the covariance matrix, as represented by the corresponding entry in the diagonal covariance matrix [12]. The number of vectors in the basis is equal to the minimum between the number of available class pattern and the dimensionality of the embedding space. Each pattern in the class can usually be described by using only the most important components. The resulting restricted basis spans a linear subspace in which the patterns of the represented class can be found. Each possible pattern $\boldsymbol{\phi}$ can be projected onto the set of principal components and can be described as the sum of its projection $\boldsymbol{\phi}_{L_i}$ plus an orthogonal residual $\boldsymbol{\phi}_{R_i}$:

$$\boldsymbol{\phi} = \boldsymbol{\phi}_{L_i} + \boldsymbol{\phi}_{R_i} + < \phi_i > \tag{50}$$

where i identifies the class and $< \phi_i >$ is the corresponding centroid. A comparison with the usual technique of computing the distance from a single pattern (e.g. the centroid) is reported in Figure 9.

An important class of objects spanning a linear space is given by the ortographic projections of rigid sets of points when looked at from different positions[1, 23]. Different objects span different 6-dimensional linear spaces. This can be used to recognize them, irrespective of their orientation in space, by computing the magnitude of the projection residual over the individual linear spaces (see Figure 9). Under perspective projection, when viewing an object from a reasonable distance, we expect that a 6-dimensional linear space can still provide a good approximation to the real manifold. Further analysis of the recognition experiments reported in [5] has shown that an HyperBF network [25] with a single unit is in fact able to learn the approximating linear space from a set of example views of different objects. The experiments used paper clips characterized by 6 feature points in the image plane, resulting in 12-dimensional vectors after perspective projection. The i-th clip was characterized by a one-unit HyperBF network:

$$C_i(\boldsymbol{\phi}) = exp(-(\boldsymbol{\phi} - \boldsymbol{t}_i)^T W_i^T W_i (\boldsymbol{\phi} - \boldsymbol{t}_i)) \tag{51}$$

where $\boldsymbol{\phi}$ is the 12-dimensional input to the network, $\boldsymbol{t}_i$ is a sample view (a prototype) of the i-th clip and $W_i^T W_i$ represents a metric. The network is trained by modifying $\boldsymbol{t}_i$ and $W_i^T W_i$ to obtain $C_i(\boldsymbol{\phi}) \approx 1$ when $\boldsymbol{\phi}$ is a view of the i-th clip and $C_i(\boldsymbol{\phi}) \approx 0$ when it is not. The effects of the resulting metric $W_i^T W_i$ on the computation of distances between different views of the clips can be seen

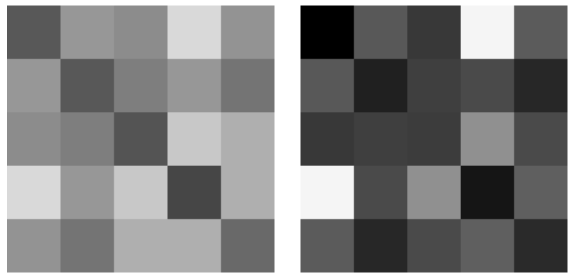

Figure 10: The square of coordinates ij represents the average value of distances of views of the i-th and j-th clip (darker values represent smaller distances). LEFT: euclidean distances of views of the different clips; RIGHT: distances computed using the learned metric $W_0^T W_0$

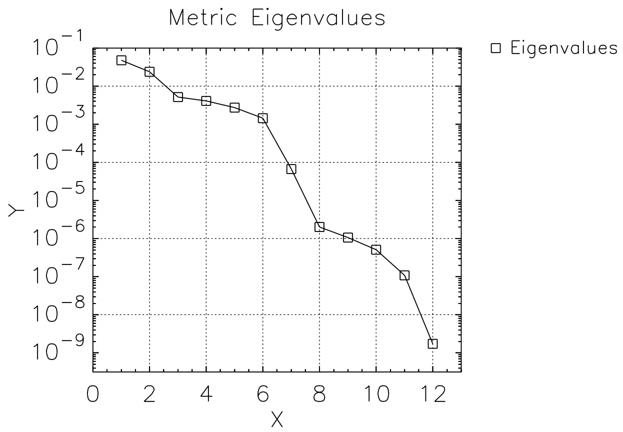

Figure 11: Eigenvalues of the learned metric matrix $\mathbf{W}^T\mathbf{W}$. Note that there is compatibility with the findings of Basri-Ullman that under ortographic projection the rank of the metric is 6.

in Figure 10. The distance computed using the learned metric is effectively the size of the projection residual. The eigenvalues of $W_i^T W_i$ (see Figure 11) are compatible with a 6-dimensional embedding of the pattern space.

If the linear subspace is the one spanned by the first k eigenvectors of the covariance matrix, the sum of the eigenvalues corresponding to the ignored components can be used as an estimate for $|\boldsymbol{\phi}_R|$ when the pattern belongs to the given class. In particular it can be used to accept or reject the pattern according to a threshold on the size of the residual

$$|\boldsymbol{\phi}_R|^2 < \delta \tag{52}$$

where

$$\delta = \beta \sum_{i>k} \lambda_i \tag{53}$$

and $\beta \geq 1$ is an heuristic factor taking into account how good an estimate $\sum_{i>k} \lambda_i$ is of the residual error. In Figure 12 we report the fraction of image pixels classified as *right eye* as a function of the threshold on the residual. The fact that the residual is small (compared to δ) does not imply that the pattern belongs to the given class. Thresholding on the residual error should then be supplemented by the use of classification techniques in each of the linear subspaces, taking into account the distributions of the patterns. If, for instance, the distribution of the points in the linear subspace is Gaussian, the parameters of the distribution can be computed and the probability of a pattern with given coordinates estimated (see Figure 13 where an example is reported). If we denote by x_i the i-th component of ϕ the following relation holds if the distribution in the feature space is Gaussian:

$$\mathcal{P} \propto \Pi_i e^{-x_i^2/(2\lambda_i)} \tag{54}$$

where n is the number of patterns used for computing the principal components. The resulting map can be used in conjunction with the distance map (see Figure 14) to establish if a pattern of the correct class is present (for a similar approach, using the distance from the centroid in the projection space [29, 30, 31]). Note that in this particular case the probability map is much more effective than the residual distance map.

It could be that a class cannot be packed tightly into a linear subspace. A possible improvement is to attempt *local expansions* [27, 22, 31]. Points can be clustered, and for each of the resulting clusters a principal component analysis can be attempted. The previous reasonings can be applied and the class is represented by a set of linear subspaces. A nice application of this approach can be found in [22] where the space spanned by faces is first clustered into views corresponding to different poses and the resulting clusters are then described by the most important principal components.

7 An experimental comparison

In order to clarify the practical relevance and the relative merits of the previous template-matching techniques, it is useful to compare them on a single task. We choose to assess the performance of the different techniques on the problem of locating eyes in frontal images of faces. This

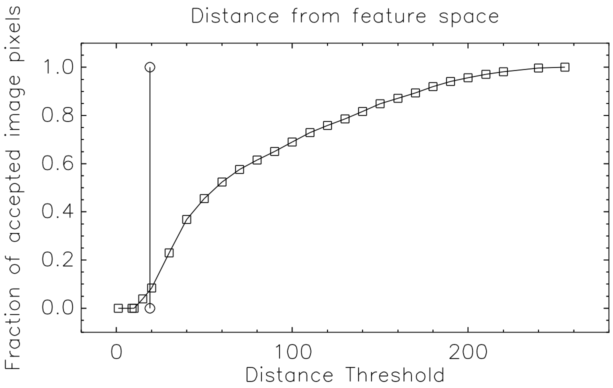

Figure 12: The fraction of image pixels classified as *right eye* as a function of the threshold on the residual. The first 10 eigenvectors from a population of 60 images were used. The image used for the plot was of a person not in the database. The vertical line represent the threshold computed by summing the residual eigenvalues. The correct eye is the only selected region for $d < 11$, the other eye being selected next.

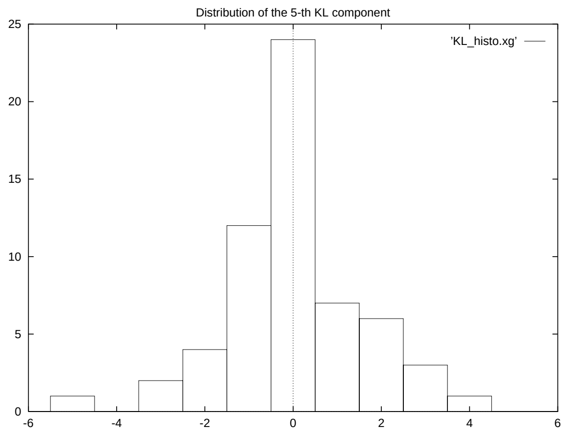

Figure 13: The distribution of the values of the 5-th principal component computed from 60 eyes images. Note the clear unimodality of the distribution which suggests the effectiveness of using a quadratic classifier in the feature subspace. The other components present a similar distribution.

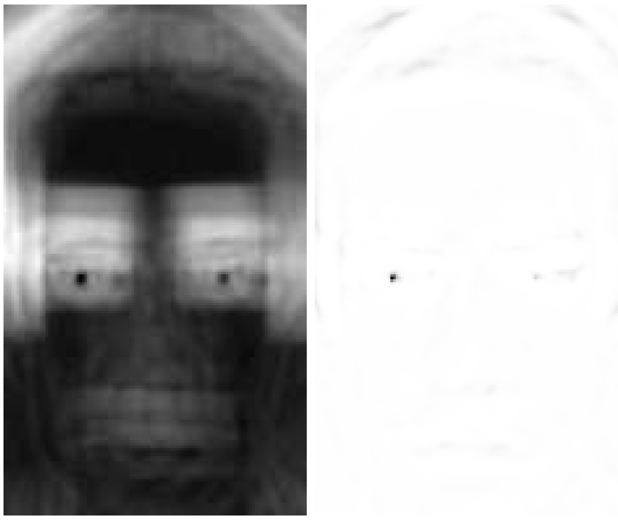

Figure 14: The map of the residual size (left) and of the projection probability (right). Note how the probability is low in regions where the reconstruction is good. The darker the value the lower the distance and the higher the probability.

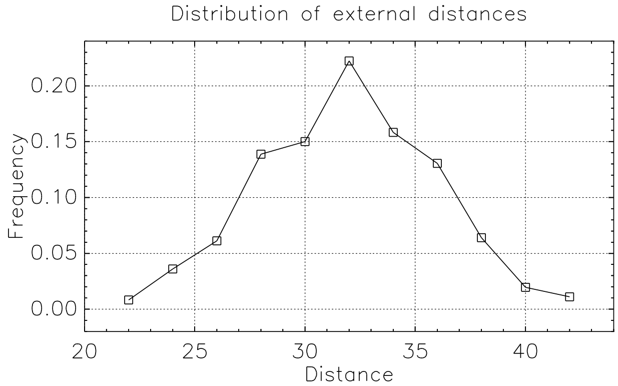

Figure 15: Distribution of the distance values orthogonal to the feature space when projecting onto the first 10 eigenvectors.

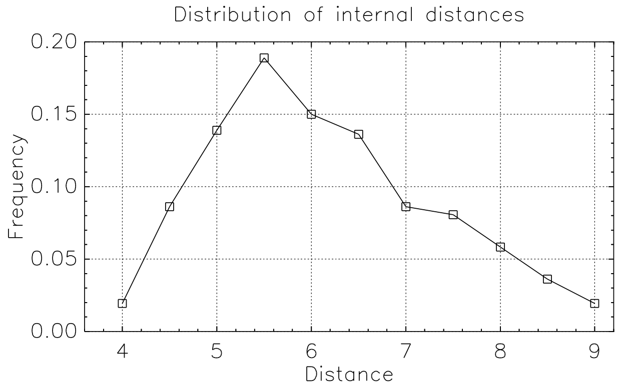

Figure 16: Distribution of the distance values within the feature space when projecting onto the first 10 eigenvectors.

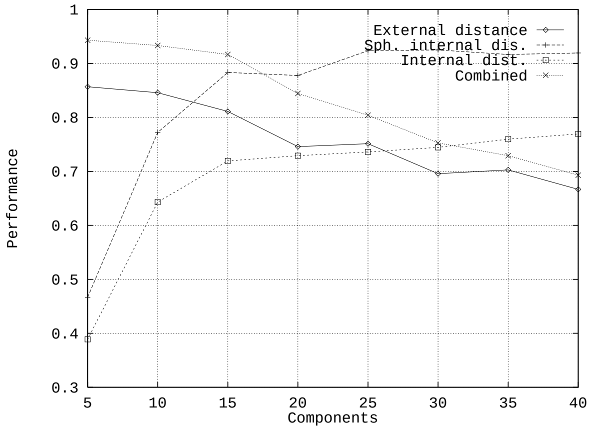

Figure 17: Performance of different strategies based on the computation of principal components. The horizontal axis reports the number of components used in the expansion, while the vertical axis reports the percentage of eyes correctly located (see text for a detailed explanation).

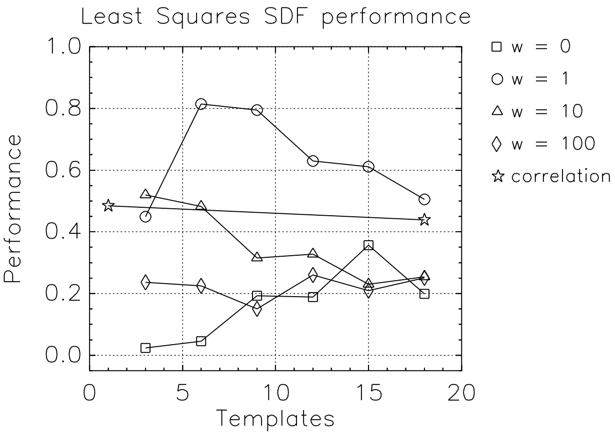

Figure 18: Performance of least squares SDF with different amount of regularizing noise. Correlation performance is also reported using the average template and the whole set of available templates. The horizontal axis reports the number of patterns used in building the filters, while the vertical axis reports the percentage of eyes correctly located (see text for a detailed explanation).

is a preliminary step for identifying the represented person by comparing his/her image to a reference database. The available database consisted of 180 images (three images, taken at different time, from sixty different people). The eyes were manually located and images normalized by fixing the position of the eyes to standard values. The resulting normalized images (with an interocular distance of 28 pixels) were used for the experiments. Three different disjoint subsets, each consisting of the images from 18 different people were used in turn for building the SDFs, lsSDF and KL expansions. Performance was then assessed on the remaining images.

For each of the compared strategies and testing images, a map was computed reporting at each pixel the absolute difference of the computed values (residual, correlation, etc.) from the required values at the pattern (i.e. eye) location (e.g. 0 for the residual, 1 for correlation). The resulting maps could then be considered as *distance* maps. For each image we masked in turn the region of the left and right eye. The unmasked eye was considered to be located correctly if the smallest *distance* value was within 8 pixels from the correct location (manually detected).

As far as the SDFs and lsSDF are concerned, a single image from the represented persons was used in building the filters, while the computation of the KL components relied on all the available images in the *training* subset. For all of the techniques the test was run on 120 images. Both left and right eyes were used in building the filters and the expansions. In order to assess the performance of the techniques, each image was used to locate both eyes by masking in turn the left and right eye region when looking for the maximum/minimum values ideally associated to the template location.

Several variants of the KL approach have been investigated using distances from and within the feature space. The *external distance* d_e is simply the error in the reconstruction of the pattern using the restricted eigenvector basis (see eqn. 50). The *internal distance* d_i is the distance computed within the linear subspace from its origin (the centroid of the patterns). The *spherical internal distance* d_{s_i} is the Mahalanobis distance in the linear subspace.

Let us assume that the orthogonal vectors defining the linear pattern subspace are known or computed reliably from a subset of the available examples. We could then estimate the Mahalanobis distance by computing the variance for each of the (uncorrelated) coordinates using all the available samples. Some of the examples could be erroneous or atypical and would probably lead to an overestimated variance. In order to overcome this potential problem, a robust estimate of the scale parameter of each coordinate was computed using the *tan-h* M-estimators introduced by Hampel [13]. Finally the *combined distance* (see also the related approach in [29, 30, 31]) was computed by the following relation:

$$d_c = \max\left(\frac{d_i}{d_{s0}}, \frac{d_e}{d_{e0}}\right) \tag{55}$$

where the normalizing factors d_{s0} and d_{e0} define the points at which the cumulative distribution of the *internal* and *external* distances reaches 99% (see Figures 15 and 16). The performance of the different variants are reported in Figure 17.

SDFs and lsSDF were built using different amounts of regularizing noise (using eqn. (40)) and of templates. The resulting performance, together with the performance of standard correlation is reported in Fig. 18. It is interesting to note the major impact of the regularizing noise on the performance of this technique. However, the bias and variance of the filter responses on the test images are not related to the filter performance.

The combined distance d_c is the best among the compared strategies. Its decline in performance with increasing dimensionality of the expansion basis is linked to the trend of the external distance performance. By using more and more eigenvectors we allow for good reconstruction of patterns different from eyes. At the same time the scaling factor computed from the distance distribution on the *training* samples becomes very small (should we use all of the computed eigenvectors the samples could be reconstructed exactly). Therefore the external distance is the (wrong) dominating factor in eqn.(55). A more sophisticated integration is presented in [31]. The performance of the template matching strategies based on KL expansions is consistently higher than the one achieved by SDFs in the reported variants. Also, expanding a pattern onto an appropriate basis seems to provide reliable template matching to patterns which span a manifold which can be approximated well (at least locally) by a linear (tangent) space [1, 24, 23, 5].

The next section will introduce non linear machinery (sigmoidal and Gaussian network) for the purposes of pattern description and classification.

8 Future Directions: Learning and SDF

The description of the advanced SDFs has shown that that they can be considered as standard SDF working on a preprocessed signal. The characteristics of the original signal and noise are used in the synthesis of the preprocessing filter to achieve optimal sharpness in the correlator response. If we look at the patterns in the transformed space, the correlator output is a weighted average of the correlation with a set of examples:

$$\phi'(x) \otimes h'(x) = \sum_{i=1,\ldots,n} b'_i \phi'(x) \otimes \phi'_i(x) \tag{56}$$

where the prime refers to the transformed space. The patterns $\boldsymbol{\phi}_i{}'$ can be randomly chosen among the available examples or selected according to particular criteria. A possible strategy is to synthesize the filter incrementally: the response of the filter on all the training images not yet used to build the filter is computed and if the worst filter response is not acceptable the corresponding image is added to the building set and a new filter is computed [7]. The construction of the filter, apart from the phase of selecting meaningful training images is linear. An improvement is expected with the introduction of nonlinearity in the filter design. We propose the use of approximation networks [25] to build general non linear filters which are able to discriminate patterns of different classes while giving the same response on patterns of the same class. These filters can be considered as a generalization of the projections Synthetic Discriminant Functions. They are built using a set of training images and a set of soft (i.e. not exactly met) constraints on the filter output.

The general structure of the network is reported in Figure 19. The units of the first level represent sigmoidal (comparison by projection) or Gaussians (comparison by distance) functions:

$$o_{1i}(\boldsymbol{\phi}) = \begin{cases} exp(-(\boldsymbol{\phi} - \boldsymbol{t}_i)^T W^T W(\boldsymbol{\phi} - \boldsymbol{t}_i)) & \text{Gaussian} \\ \sigma(\boldsymbol{\phi} \cdot \boldsymbol{t}_i + \beta) & \text{sigmoidal} \end{cases} \quad (57)$$

In both cases, the system is able to mask regions of the templates which are not useful for obtaining the required output values[2]. The first level of the network can be seen as computing some "optimal templates" against which the input signal is to be compared. The output of the second level is computed as:

$$\sum_j b_j o_{2j}(\boldsymbol{o}_1) \quad (58)$$

and the function implemented by unit o_{2j} can be of the Gaussian or sigmoidal type independently from the choice of the first layer units. The second level computes a non linear mapping of the projections (or distances) of the signal by minimizing the square error of the network on the mapping constraints (soft, as they are not met exactly). In some sense, the network triangulates the position of the input signal in pattern space using the distances from automatically selected reference templates. The resulting networks have a very high number of free parameters and their training presents difficulties. Among them two are of particular concern: overfitting and training time in a high dimensional space. A way of coping with the first one is that of cross-validation [12]: the network undergoes training as long as its performance on a test set improves. We propose to reduce the effects of high dimensionality by using a hierarchy of networks of similar structure but working at increasing resolution. The network with the lowest resolution is trained first and extensively. The next network in the hierarchy is then initialized by suitably mapping the parameters of the previous one. Note that only the first level needs to be modified structurally. A reduced training time is expected. The procedure is iterated at all the levels of the hierarchy. A side effect of the hierarchical training is to provide fully trained networks for different resolutions, enabling a hierarchical approach to template matching. The preprocessing stage of the network is the one computed for the synthesis of the ASDF. The *optimal templates* used by the first layer of the network can also be initialized using the building patterns of the linear filter.

[2] This is achieved during the training phase by modification of the entries of the matrix W, if Gaussians are used, or t_i if sigmoids are used. Relatively small values give low weight to the differences of the corresponding coordinates, thereby making the system output weakly dependent on them.

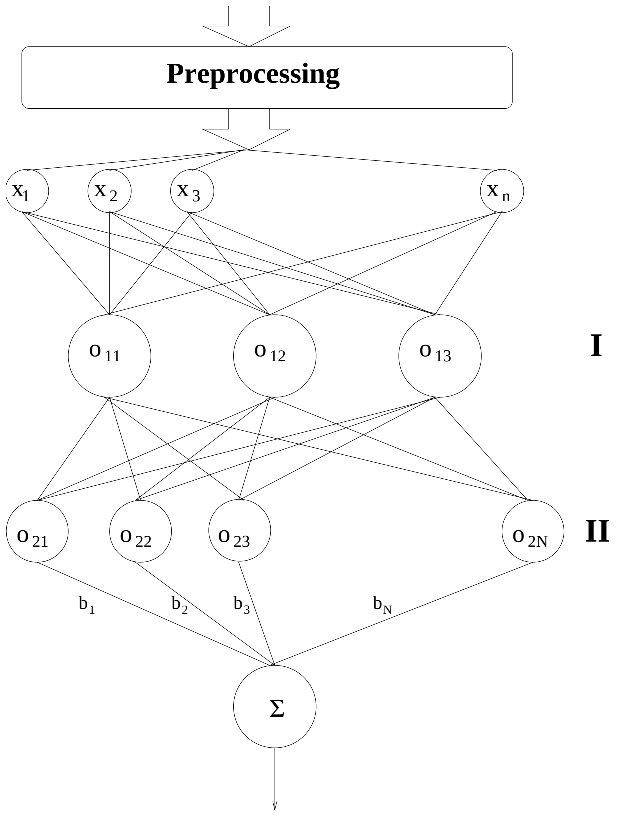

Figure 19: An approximation network for template matching. The preprocessing stage applies simple transformation to the input pattern (e.g. to emphasize high frequency components).

9 Conclusions

Several approaches to template matching have been reviewed and compared on a common task. A new variant of Synthetic Discriminant Functions, based on least square estimation, was introduced. Several template matching techniques based on the expansion of patterns on principal components have been reviewed. A simple way of integrating internal/external distances within/from a linear feature space was also proposed. Several of the techniques mentioned in the paper have been compared on a common task: locating eyes in frontal images of different people. The techniques based on pattern expansion provide superior performance, at least in the particular task considered. Finally, a two layer approximation network has been proposed to generalize the structure of SDF to a nonlinear filter. Future work will explore the advantages and difficulties of the introduction of nonlinearity.

References

[1] R. Basri and S. Ullman. Recognition by linear combinations of models. Technical report, The Weizmann Institute of Science, 1989.

[2] J. Ben-Arie and K. R. Rao. A Novel Approach for Template Matching by Nonorthogonal Image Ex-

pansion. *IEEE Transactions on Circuits and Systems for Video Technology*, 3(1):71–84, 1993.

[3] J. Ben-Arie and K. R. Rao. On the Recognition of Occluded Shapes and Generic Faces Using Multiple-Template Expansion Matching. In *CVPR '93*, pages 214–219, 1993.

[4] R. Brunelli and S. Messelodi. Robust Estimation of Correlation: an Application to Computer Vision. Technical Report 9310-05, I.R.S.T, 1993. to appear on *Pattern Recognition.*

[5] R. Brunelli and T. Poggio. HyperBF Networks for real object recognition. In John Mylopoulos and Ray Reiter, editors, *Proc. 12th IJCAI, Sidney*, pages 1278–1284. Morgan-Kauffman, 1991.

[6] D. Casasent and Wen-Thong Chang. Correlation synthetic discriminant functions. *Applied Optics*, 25(14):2343–2350, 1986.

[7] D. Casasent and G. Ravichandran. Advanced distortion-invariant minimum average corelation energy (MACE) filters. *Applied Optics*, 31(8):1109–1116, 1992.

[8] D. Casasent, G. Ravichandran, and S. Bollapragada. Gaussian-minimum average correlation energy filters. *Applied Optics*, 30(35):5176–5181, 1991.

[9] David Casasent. Unified synthetic discriminant function computational formulation. *Applied Optics*, 23(10):1620–1627, 1984.

[10] H. J. Caulfield and W. T. Maloney. *Applied Optics*, 8:2354, 1969.

[11] H. J. Caulfield and M. H. Weinberg. Computer Recognition of 2-D Patterns using Generalized Matched Filters. *Applied Optics*, 21:1699, 1982.

[12] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

[13] F. R. Hampel, P. J. Rousseeuw, E. M. Ronchetti, and W. A. Stahel. *Robust statistics: the approach based on influence functions.* John Wiley & Sons, 1986.

[14] R. R. Kallman. Construction of Low Noise Optical Correlation Filters. *Applied Optics*, 25:1032–1033, 1986.

[15] B. V. K. Vijaya Kumar. Minimum-variance synthetic discriminant functions. *Journal of the Optical Society of America, A*, 3(10):1579–1584, 1986.

[16] B. V. K. Vijaya Kumar and J. D. Brasher. Relationship between maximizing the signal-to-noise ratio and minimizing the classification error probability for correlation filters. *Optics Letters*, 17(13):940–942, 1992.

[17] B. V. K. Vijaya Kumar, D. Casasent, and H. Murakami. Principal-component imagery for statistical pattern recognition correlators. *Optical Engineering*, 21(1):43–47, 1982.

[18] B. V. K. Vijaya Kumar, A. Mahalanobis, S. Song, S. R. F. Sims, and J. F. Epperson. Minimum squared error synthetic discriminant functions. *Optical Engineering*, 31(5):915–922, 1992.

[19] A. Mahalanobis and D. Casasent. Performance evaluation of minimum average correlation energy filters. *Applied Optics*, 30(5):561–572, 1991.

[20] Abhijit Mahalanobis, B. V. K. Vijaya Kumar, and David Casasent. Minimum average correlation energy filters. *Applied Optics*, 26(17):3633–3640, 1987.

[21] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, 1989.

[22] A. Pentland, B. Moghaddam, T. Starner, O. Oliyide, and M. Turk. View-Based and Modular Eigenspaces for Face Recognition. Technical Report 245, M.I.T Media Lab, 1993.

[23] T. Poggio. 3D Object Recognition: on a result by Basri and Ullman. Technical Report 9005-03, I.R.S.T, 1990.

[24] T. Poggio and S. Edelman. A Network that Learns to Recognize Three-Dimensional objects. *Nature*, 343(6225):1–3, 1990.

[25] T. Poggio and F. Girosi. Networks for Approximation and Learning. In *Proc. of the IEEE, Vol. 78*, pages 1481–1497, 1990.

[26] G. Ravichandran and D. Casasent. Minimum noise and correlation energy optical correlation filter. *Applied Optics*, 31(11):1823–1833, 1992.

[27] P. Y. Simard, Y. LeCun, and J. Denker. Efficient pattern recognition using a new transformation distance. In *Advances in Neural Information Processing Systems*, pages 50–58. Morgan Kaufman, San Mateo, 1993.

[28] S. I. Sudharsanan, A. Mahalanobis, and M. K. Sundareshan. Unified framework for the synthesis of synthetic discriminant functions with reduced noise variance and sharp correlation structure. *Optical Engineering*, 29(9):1021–1028, 1990.

[29] Kah-Kay Sung. Network Learning for Automatic Image Screening and Visual Inspection. In *Proceedings of the CBCL Learning Day*, at the American Academy of Arts and Sciences, Cambridge, Massachusetts, January 1994. CBCL at MIT.

[30] Kah-Kay Sung. personal communication. 1994.

[31] Kah-Kay Sung and Tomaso Poggio. Example-based Learning for View-based Human Face Detection. In *Proceedings Image Understanding Workshop*, 1994.

[32] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.